December 9, 2019

LexRank: Graph-based Lexical Centrality as Salience in Text Summarization

1615052 山元　悠貴

富山県立大学　情報基盤工学講座

1. はじめに
2. **introduction**
3. 実験
4. おわりに

December 9, 2019

1. はじめに

目的

グラフに基づいた自然言語処理におけるテキスト単位の相対的重要度を算出する方法を導入し、テキスト要約 **(TS)** 問題でこの手法をテストするグラフベースの手法は文章をグラフ構造で表現し各ノード (=文や単語) の関係性を元に要約を作成するという手法である

抽出型要約

抽出型要約 **(extractive TS)** は文書または文書セット内のもっとも重要だと思われる文を抽出して要約を作成することである
要点は特定の重要語の存在または疑重心文との類似性の観点で定義される

1. はじめに

目的

グラフベースの手法を使った代表的なアルゴリズムは **TextRank** というものがある **TextRank** は **Google** 検索の基礎となっている **PageRank** という手法を文章に応用した手法である

抽出型要約

抽出型要約 **(extractive TS)** は文書または文書セット内のもっとも重要だと思われる文を抽出して要約を作成することである
要点は特定の重要語の存在または疑重心文との類似性の観点で定義される

2. 提案手法

自然言語処理

自然言語処理 (NLP) での多くの問題, 解析, 語義の曖昧性解消および自動言い換えは統計的手法の導入によって大幅に恩恵を受けている. NLP に対するグラフに基づいた手法も, 単語クラスタリングや前置詞フレーズ添付で多くの関心を集めている.

2. 提案手法

pagerank

ページ間のリンクは行列で表現することができ, 各ページにおけるリンクの総数で割ることでこの行列はユーザーがそのページからリンクが張られた別のページに遷移する確率の行列になる.

Page	Link to
A	B, C, D, E
B	C, E
C	D
D	B, D, E
E	A, C

	A	B	C	D	E
A		1	1	1	1
B			1		1
C				1	
D		1		1	1
E	1		1		

Transition Matrix

	A	B	C	D	E
A		1/4	1/4	1/4	1/4
B			1/2		1/2
C				1	
D		1/3		1/3	1/3
E	1/2		1/2		

Transition Probability Matrix (=M)

Page Surfer	
A	0.2
B	0.1
C	0.3
D	0.1
E	0.3
	1

Page Surfer State (=P)

MP = P when t -> ∞

Solution

1. Solve Eigenvalue Problem of **MP = P**.
2. Repeat the transition until convergence(**MP - P < threshold**).

$$P_i' = (1-d) + d * M_i^T P_i$$

$$\sum (P_i' - P_i) < threshold$$

*The page surfers randomly click the page with a probability of **1-d**. (d = usually 0.85)*

2. 分野

提案手法

文章のグラフ表現における固有ベクトル中心性をコンセプトとして文の重要度を計算する新しいアプローチである LexRank を考察する．これをテキスト要約に用いる．
テキスト要約はユーザに有用な情報を提供するために特定のテキストの圧縮版を自動で生成する処理である．ここでの目的は不特定なトピックについて複数のドキュメントの要約を生成することである．
本稿ではクラスタに含まれる各文の中心性を評価し要約に含める
もっとも重要な文を抽出する

$$\mathrm{idf}_i = \log\left(\frac{N}{n_i}\right)$$

N ： 文書集合に含まれる文書の総数
n_i ： 単語 i が出現する文書の数

3. 手順

ステップ 1

クラスタの重心は $tf \times idf$ スコアが事前定義されているしきい値を上回る単語からなる疑似文書である.
ここで tf はクラスタ内の単語の頻度である.

$$\mathrm{idf}_i = \log\left(\frac{N}{n_i}\right)$$

N : 文書集合に含まれる文書の総数
n_i : 単語 i が出現する文書の数

3. 手順

提案手法

2 つの文の類似性は 2 つの対応するベクトル間のコサインによって定義される

$$\text{idf\_modified\_cosine}(x, y) = \frac{\sum_{w \in x,y} \text{tf}_{w,x}\text{tf}_{w,y}(\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x}(\text{tf}_{x_i,x}\text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(\text{tf}_{y_i,y}\text{idf}_{y_i})^2}}$$

この式から 2 つの文のコサイン類似度を求めることができる
ここからコサイン類似度行列および対応するグラフ表現を用いていくつかの文中心性の算出方法を考察する.

3. 手順

実験

関連文書クラスタ内では文の多くが同一のトピックに関するものであるため、文の多くはお互いにやや似ている
今回, 11 の文に対してそれぞれのコサイン類似度を算出して隣接行列にした

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18

グラフ化

Figure1 の行列に含まれるクラスタの重み付きコサイン類似度をグラフ化する

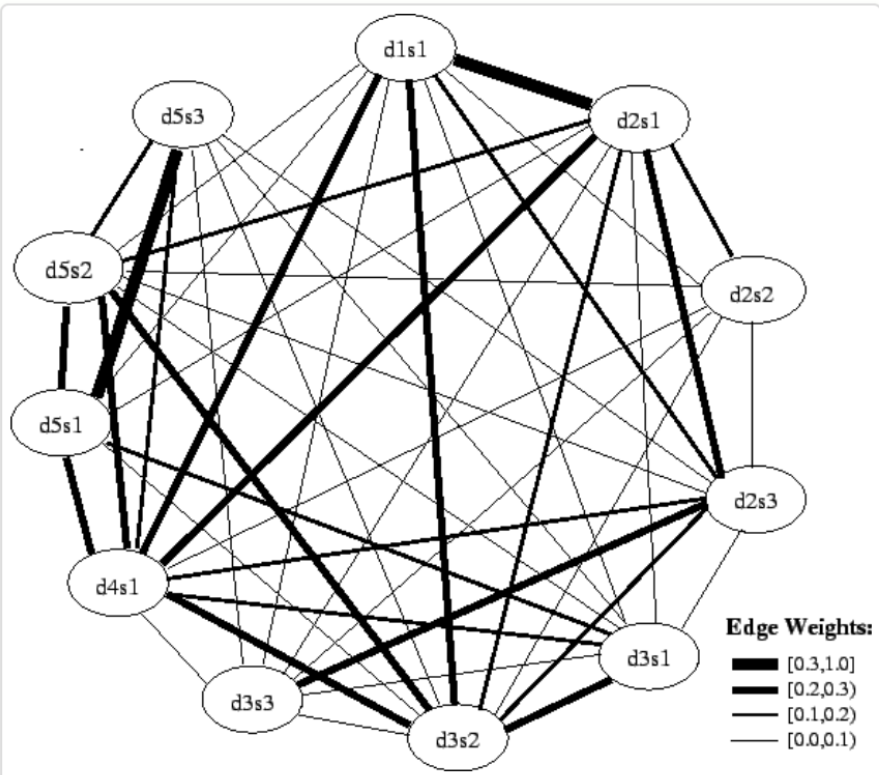

Figure 2: Figure 1 に含まれるクラスタの重み付きコサイン類似度グラフ。 (訳注: 連続 LexRank は完全グラフになるため d1s1 d5s3 間の Edge が足りない)

3. 手順

ステップ 3

無向 グラフとしてみなすために、しきい値を定義することによって行列内のいくつかの低い値を排除する
Figure 3 は、**Figure 1** における類似度がそれぞれ **0.1 0.2, 0.3** を超える文ペアが互いに類似していると仮定して導かれた隣接行列に対応するグラフを示している

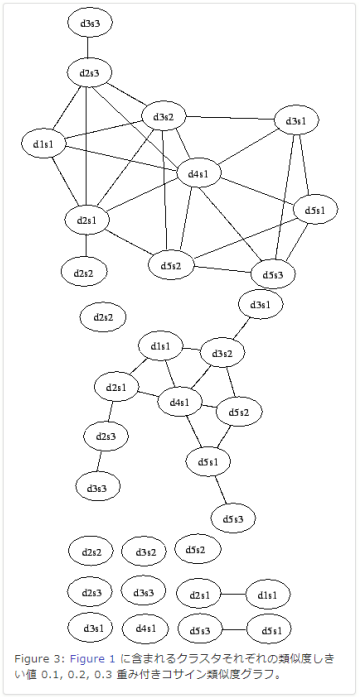

Figure 3: Figure 1 に含まれるクラスタそれぞれの類似度しきい値 0.1, 0.2, 0.3 重み付きコサイン類似度グラフ。

3. 手順

中心的な文を探す

コサイン類似度がしきい値以上のグラフのエッジの数が多いノードを探す
低すぎるしきい値は誤った類似性を考慮させる可能性があるが，高すぎるしきい値もクラスタ内の類似関係の多くを失う可能性がある

Table 1: Figure 3 中のグラフの次数中心スコア。しきい値 0.1 と 0.2 で d4s1 が最も中心的な文。

ID	Degree (0.1)	Degree (0.2)	Degree (0.3)
d1s1	5	4	2
d2s1	7	4	2
d2s2	2	1	1
d2s3	6	3	1
d3s1	5	2	1
d3s2	7	5	1
d3s3	2	2	1
d4s1	9	6	1
d5s1	5	4	2
d5s2	6	4	1
d5s3	5	2	2

3. 手順

次数の中心は，いくつかの望ましくない文章が互いに投票してその中心性を上げている場合，要約の質に悪影響を及ぼすことがある
投票がどこから来ているかを考慮し、各投票で重み付けする際に投票ノードの中心性を考慮することによって避けることができる
中心値を有するすべてのノードを考慮し、この中心性をその隣接ノードに分配する．
これを定式化すると

$$p(u) = \sum_{v \in \mathrm{adj}[u]} \frac{p(v)}{\deg(v)} \tag{1}$$

ここで $p(u)$ はノード u の中心性、$\mathrm{adj}(u)$ はノード u に隣接しているノードの集合、$\deg(u)$ はノード u の次数。同様に式 (1) を行列で記述することもできる。

$$\boldsymbol{p} = \boldsymbol{B}^T \boldsymbol{p}$$

ここで行列 $\boldsymbol{B}$ は、類似度グラフの隣接行列の各要素を対応する行の和で除算することで得られる。

$$\boldsymbol{B}(i,j) = \frac{\boldsymbol{A}(i,j)}{\sum_k \boldsymbol{A}(i,k)}$$

4. 結果

類似行列 B が確率行列の特性を満たすことから、我々はこれをマルコフ連鎖とみなすことができる
グラフ内の任意のノードにジャンプする均一な確率を割り当てると以下の式を導ける

これは PageRank として知られている。

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in \mathrm{adj}[u]} \frac{p(v)}{\deg(v)}$$

ID	LR (0.1)	LR (0.2)	LR (0.3)	Centroid
d1s1	0.6007	0.6944	1.0000	0.7209
d2s1	0.8466	0.7317	1.0000	0.7249
d2s2	0.3491	0.6773	1.0000	0.1356
d2s3	0.7520	0.6550	1.0000	0.5694
d3s1	0.5907	0.4344	1.0000	0.6331
d3s2	0.7993	0.8718	1.0000	0.7972
d3s3	0.3548	0.4993	1.0000	0.3328
d4s1	1.0000	1.0000	1.0000	0.9414
d5s1	0.5921	0.7399	1.0000	0.9580
d5s2	0.6910	0.6967	1.0000	1.0000
d5s3	0.5921	0.4501	1.0000	0.7902

5. おわりに

まとめ

1. 自然言語処理の分野で **PageRank, LexRank** のシステムついて述べた.
2. グラフに基づいた中心性は重心に比べていくつかの利点がある
3. ノードの次数はその文が他の文とどのくらいの共通情報を有するかの指標である
4. 不自然に高い **IDF** スコアがトピックに無関係な文のスコアを上げることを防ぐ

課題

1. 類似リンクの強さに得られたコサイン値を直接使用することでより高密度で重みづけされたグラフが得られる